心理科学进展 2021, Vol. 29, No. 9, 1696–1710
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2021.01696

·研究方法(Research Method)·

心理与教育测验中异常作答处理的新技术: 混合模型方法*

刘 玥[1] 刘红云[2,3]

([1]四川师范大学脑与心理科学研究院, 成都 610066)
([2]应用实验心理北京市重点实验室) ([3]北京师范大学心理学部, 北京 100875)

摘 要 混合模型方法(Mixture Model Method)是近年来提出的, 对心理与教育测验中的异常作答进行处理的方法。与反应时阈值法, 反应时残差法等传统方法相比, 混合模型方法可以同时完成异常作答的识别和模型参数估计, 并且, 在数据污染严重的情况下仍具有较好的表现。该方法的原理为根据正常作答和异常作答的特点, 针对分类潜变量的不同类别, 在作答反应和反应时部分建立不同的模型, 从而实现对分类潜变量(即作答层面的分类), 以及模型中其他题目和被试参数的估计。文章详细介绍了目前提出的几种混合模型方法, 并将其与传统方法比较分析。未来研究可在模型前提假设违背, 含有多种异常作答等情况下探索混合模型方法的稳健性和适用性, 通过固定部分题目参数, 增加选择流程等方式提高混合模型方法的使用效率。

关键词 异常作答, 反应时, 阈值, 残差法, 混合模型
分类号 B841

1 引言

在使用心理与教育测验对学生的人格、技能和能力等潜在特质进行测量时, 最主要的目的是基于测验信息得到学生潜在特质的有效估计值。然而, 在实际中, 学生完成测验时往往不可避免地因为异常作答的出现带来一些与测验结构无关的污染。异常作答根据其原因可以分为不努力作答(non-effortful responses)、对题目有预了解的作答(preknowledge, Qian et al., 2016; Sinharay & Johnson, 2019; Wang, Xu, Shang, & Kuncel, 2018)和作弊等。不同原因的异常作答可能有不同的表现。例如, 不努力作答可能表现为忽略题目、加速作答(speededness, Hong & Cheng, 2019b; Shao et al., 2016; Yu & Cheng, 2019)、快速猜测作答(rapid-guessing, Wise, 2015, 2017)等。

异常作答在心理和教育测验中非常常见。例如, 在明尼苏达多项人格测验(Minnesota Multiphasic Personality Inventory, MMPI; Baer et al., 1997; Berry et al., 1992)的一些测试中, 有超过 50%的被试快速猜测作答的题目数在 1 道以上。根据美国国家自然科学基金(www.nsf.gov/statistics/seind93/chap1/doc/1s193.htm)统计, 接近一半(45%)的 12 年级学生报告他们在国家教育进展评估(National Assessment of Educational Progress, NAEP)的数学测验表现不如他们在学校测验上努力。Bridgeman 和 Cline (2004)发现有几乎一半的被试在 CAT-GRE (基于计算机的美国研究生入学考试)的最后 6 道题目上存在加速作答行为。

总的来说, 异常作答具有提供的心理测量学信息少的特征(Wise, 2015, 2017)。因此, 如果在测验中出现了异常作答行为, 那么传统测量模型就不能恰当处理, 造成有偏差的估计结果。首先, 很多情况下被试的能力估计值会出现偏差(Rios et al., 2017; Wise, 2015; Wise & DeMars, 2006; Wise & Kingsbury, 2016), 进而造成群组分数的差异

收稿日期: 2020-10-23
* 国家自然科学基金项目(32071091)。
通信作者: 刘红云, E-mail: hyliu@bnu.edu.cn

(Borghans & Schils, 2012)。其次，题目参数估计值的偏差会增大(Schnipke & Scrams, 2002; Wise & DeMars, 2006)。第三，如果不同子群体中异常作答的比例不同，这种差异还可能导致项目功能差异，或者对不同子群体测验表现的比较存在偏差(Setzer et al., 2013; Wise & DeMars, 2010)。第四，测验的信息量、信度会出现偏差(Wise & DeMars, 2006)。例如，原有的分析方法将无效的异常作答视为有效，可能会高估信度。第五，测验所测量的结构也可能会发生变化，会聚效度出现偏差(Weirich et al., 2017; Wise & DeMars, 2006)。最后，与测验有关的预测变量和结果变量之间的关系，假设检验得到的结论等，都可能会出现偏差(Clark et al., 2003)。综上，异常作答不仅会造成被试潜在特质估计值的偏差，也会降低测验质量相关指标的准确性，对标定测验题目参数、开发测验等造成严重影响。因此在测验的数据分析中，有必要通过科学的方法，处理异常作答，减小其不利影响，得到更准确的参数估计结果。

异常作答的处理主要分为识别并降低权重，在模型中处理两种思路(Morgenthaler, 2007)。异常作答传统的处理方式主要是识别并降低权重，它是指在数据清理时首先识别异常作答，再在数据分析时降低异常作答在样本中的权重(Ranger et al., 2019; Rios et al., 2017)。一种降低权重的处理方式是采用稳健的估计方法(Hong & Cheng, 2019a)。而降低权重中最极端的方式是替换为缺失。在异常作答比例不太大的情况下，这种方式得到的参数估计结果是可以接受的(e.g., Custer et al., 2012; Köhler et al., 2017; Rose, 2013)。然而，这类方法主要存在两个问题。一是在识别阶段，关于如何确定有效、可信的阈值，往往存在较大争议。二是在降低权重阶段，当异常作答与所测量的潜在特质相关时(Wise, 2017)，如果异常作答的比例较大，那么这种方式得到的参数估计值也是有偏的。为了解决这一问题，近年来一些研究者提出了在模型中处理的方法。该方法主要指使用混合模型对整体数据建模，正常作答和异常作答的数据分别采用不同的模型拟合(Meyer, 2010; Molenaar et al., 2018; Wang & Xu, 2015; Wang, Xu, & Shang, 2018; Wise & DeMars, 2006)。这种方法的优势在于能够一次性解决异常作答识别和参数估计的问题。并且，即使异常作答与所测量的潜在特质有关(即类似于非随机缺失)，无法简单采用降低权重的方式处理，很多研究证明基于模型的方法也能够较好地处理这种数据(Pohl et al., 2012; Rose et al., 2017)。

混合模型在识别异常作答上的应用最早可以追溯到 Schnipke 和 Scrams (1997)使用对数正态混合模型拟合反应时数据，以区分努力作答和不努力作答的被试。他们假设，如果每名被试在每道题目上的作答行为都可以被分为认真作答(solution behavior)或不努力作答，并且，这两种作答行为有不同的反应时分布。那么，每道题目上的反应时分布就是两种行为反应时的混合分布。即，可以使用二元正态分布的混合模型拟合反应时。后来，Bolt 等人(2002)又提出使用混合 Rasch 模型从作答反应方面区分含加速行为和不含加速行为的被试。该模型假设在测验最末的题目上，含加速行为的潜类别估计得到的难度参数高于不含加速行为的潜类别估计结果。因此，可以使用贝叶斯估计的方法定义待估参数的先验分布进行估计。最初的混合模型方法有两个方面的缺陷。一是仅针对反应时或者作答反应建立混合模型，没有同时利用两方面信息。根据不努力作答具有反应时短、正确率低的特点，或者对题目有预了解的作答具有反应时短、正确率高的特点，如果能够同时基于反应时和作答反应的信息建立混合模型，势必能够更精准地侦查这些类型的异常作答，提高模型参数估计的准确性。二是混合模型中的类别潜变量是针对被试的，只能完成被试层面的识别。但是在整个测验中，被试正常作答和异常作答的状态可以来回转换(Wang & Xu, 2015; Wise, 2015, 2017)。即使侦别为异常作答的被试，也可能在部分题目上正常作答，反之，判断为正常作答的被试，也可能在极少题目上异常作答。因此，为了最大程度保留有效数据并提高模型参数估计精度，混合模型应能够实现作答层面的分类(Patton et al., 2019; Yu & Cheng, 2019)。

为克服以往混合模型的缺陷，近年来发展起来的用于处理异常作答的混合模型不仅同时利用了反应时和作答反应的信息建模，也可以实现作答层面的识别(Pokropek, 2016; Wang & Xu, 2015)。然而，这些方法虽然得到了国外研究者的广泛关注，但仍处于方法的提出阶段，缺乏对于方法适用性的模拟研究或应用研究。而国内学者对于心

理与教育测验中的异常值多采用拓展为四参数 IRT (item response theory, IRT)模型(如猜测现象, 见简小珠 等, 2010), 或利用个人拟合指标识别(例如作弊, 见黄美薇 等, 2020)等方式处理。鲜有研究者采用混合模型的方式处理数据中的异常作答。因此, 本文旨在通过详细介绍基于混合模型处理异常作答的方法, 并与其他识别方法进行对比, 总结并归纳其局限性及未来研究方向, 以促进该方法在国内理论研究和实证应用的发展。

本文首先介绍心理与教育测验中异常作答的两类传统识别方法: 反应时阈值法和反应时残差法。之后详细综述基于混合模型处理异常作答的方法, 及每种方法的优缺点。再综合比较这几类方法在处理异常作答中的特点、优劣及使用时的注意事项。最后, 分析混合模型方法可以改进的方面, 并指明未来研究方向。

1.1 反应时阈值法

反应时阈值法(response time threshold method)所基于的原理是, 如果一些被试在作答某道题目时, 反应时明显小于正常被试读题、理解和作答所需要的时间(Michaelides et al., 2020; Wise, 2017)。那么可以推断这些被试在这道题目上为异常作答。这类异常作答(如加速作答、快速猜测作答等, 以下简称“快速异常作答”)具有反应时短, 提供的心理测量学信息少两个方面的特征(Wise, 2015, 2017)。因此, 对于每道题目可以确定一个反应时阈值 T_j (j 表示题目), 代表正常作答和快速异常作答行为的界限。如果被试在题目上的反应时大于阈值, 则为正常作答, 反之则为快速异常作答。

反应时阈值法中最简单的方法是统一阈值法(Kong et al., 2007)。它是指基于对题目的先验研究, 给所有题目确定统一的反应时阈值(如, 3~5 秒)。由于需要较长时间读题的题目理应有更长的阈值, 统一阈值的设定显然不合理, 因此一些学者提出了根据题目特征求阈值法(Kong et al., 2007; Silm et al., 2013)。Schnipke 和 Scrams (1997, 2002)基于大量观察发现, 包含快速异常作答的反应时分布呈现双峰分布的特点: 第一个峰值频数较小, 反应时很短, 表示快速异常作答。第二个峰值频数较大, 反应时较长, 表示正常作答。双峰分布交点求阈值法将两个分布交点所对应的反应时作为阈值。Wise 和 Ma (2012)通过大量观察发现, 当反应时超过一个固定的时间点之后, 作答正确率会从随机水平开始显著升高。这个固定的时间点就标志着正常作答和低正确率快速异常作答(例如快速猜测作答等)的分界点, 它大概等于每道题目平均反应时的 10% (同时不超过 10 秒)。常模阈值法将这个分界点作为反应时阈值。基于信息求阈值法假设, 随着反应时增加, 题目作答正确率和整个测验表现的平均正确率的相关表现出从无信息(低相关)到有信息(高相关)的转换, 发生这种转变的点(即题目得分和总分的相关为 0.2)可以作为阈值(Wise, 2019)。条件分布法是一种针对选择题的结合了反应时和正确率的求阈值方法(Ma et al., 2011; Guo et al., 2016)。它的原理是找到作答正确率等于随机水平时所对应的反应时, 作为划分两种作答行为的反应时阈值。

反应时阈值法大多基于快速异常作答的特点提出, 较简单、直接, 易于理解。并且, 在大部分应用研究中取得了较好的效果(Kong et al., 2007)。但是每种方法仍存在一定的局限性。首先, 统一阈值法尽管最简单, 但是由于不同题目特征不同, 所需读题和扫描的时间也不一定相同(Yan & Tourangeau, 2008), 对所有题目使用相同阈值显然不合理。为改进这一不足, 根据题目特征求阈值法基于题目特征设置阈值。但是要使用哪些特征确定阈值, 如何根据这些特征确定阈值也没有普遍认可的结论。其次, 双峰分布交点求阈值法最主要的问题是, 实践中有很多情况下反应时不是双峰分布。例如, 当正常作答行为所需反应时本身就很短时(Wise, 2017, 2019), 两种作答反应时的分布会交叉重叠, 反应时就不一定是双峰分布。基于信息求阈值法和常模阈值法虽然能够在双峰分布不存在的条件下应用。但是, 当题目整体区分度较低, 或者两种作答的正确率相差不大时, 基于信息求阈值法的结果不够准确。而常模阈值法仅通过经验观察提出, 其推广性仍需要经过模拟和实证研究检验。最后, 条件分布法虽然能够有效区分正常作答和低正确率快速异常作答(Guo et al., 2016; Lee & Jia, 2014), 但是, 这个方法在应用方面存在三个问题。一是由于必须已知随机水平的正确率, 因此一般只适用于单项选择题。二是需要通过观察每道题目上作答反应和反应时的分布找出阈值, 很难大批量自动化地应用于大规模测验。三是实际中存在大量累积正确率曲线与随机水平没有交点的情况, 这种情况下如

何确定阈值仍没有统一有效的结论。

1.2 反应时残差法

反应时残差法(response time residual method)将反应时模型与数据拟合, 并基于模型参数计算反应时残差或期望分布, 将实际反应时残差(或反应时)与其理论分布比较, 以识别反应时异常短的快速异常作答。目前所提出的反应时残差法主要包括基于 van der Linden (2006)的反应时模型的标准化反应时残差法(Qian et al., 2016)和基于 van der Linden (2007)的多层模型的贝叶斯残差法(van der Linden & Guo, 2008)。两种方法的区别在于, 标准化反应时残差法是借助标准化反应时残差符合标准正态分布进行判断, 仅利用了反应时信息。而贝叶斯残差法将实际的作答反应和反应时与多层模型拟合, 然后将反应时观测值与其后验预测密度比较做出判断, 同时利用了作答反应和反应时的信息。

反应时残差法的优势在于背后有特定的理论模型(分布), 不需要通过观察设定阈值, 也不存在无法找到阈值的特例, 可以大批量应用。但是, 这类方法所面临的最大问题在于, 高比例的快速异常作答会导致参数估计结果的偏差, 进而造成标准化反应时残差或反应时后验预测密度的偏差, 难以得到准确的识别结果。例如, Wang, Xu, Shang 和 Kuncel (2018)研究发现, 随着快速异常作答比例增加, 贝叶斯残差法表现显著变差。即使快速异常作答数据基于残差法假设生成, 当个人快速异常作答的比例产生于 $U(0.5, 0.75)$的均匀分布时, 贝叶斯残差法的正确识别率只有 0.301。

2 混合模型法

与识别并降低权重的两阶段方法相比, 混合模型法(mixture model method)能够一次性解决异常作答识别及参数估计的问题。并且, 贝叶斯框架下的马尔科夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC)算法的发展, 较好地解决了这类模型参数估计的问题。因此近年来不断有研究者提出使用混合模型处理测验中的异常作答。以下将这些方法分为两类进行介绍。第一类方法使用反应时来预测每个作答所属的潜类别, 第二类方法则直接将含有作答反应和反应时的多层模型拓展为混合模型, 同时估计得到各题目、被试参数和类别潜变量的估计值。

2.1 使用反应时预测类别的混合模型

2.1.1 等级分组的反应时模型

Pokropek (2016)借用等级分组模型的思想, 将反应时信息与 IRT 模型结合, 提出了等级分组的反应时模型, 专门用于识别快速猜测作答。

假设正常作答用 Rasch 模型拟合, 快速猜测作答的答对概率设为 1。答对概率可以表示为

$$P\left(Y_{ij}=1|C_{ij}=1,2\right)=\pi_{ij|z}+\left(1-\pi_{ij|z}\right)\frac{\exp(\theta_j-\beta_i)}{1+\exp(\theta_j-\beta_i)} \quad (1)$$

其中, Y_{ij} 表示被试 j 在题目 i 上的作答, C_{ij} 表示分组(1 表示猜测组, 2 表示正常组), $\pi_{ij|z}$ 表示基于协变量 Z (反应时), 将被试 j 在题目 i 上的作答分到组 1 的概率, $1-\pi_{ij|z}$ 表示分到组 2 的概率, $(exp(\theta_j-\beta_i))/(1+exp(\theta_j-\beta_i))$ 是 Rasch 模型, 其中 θ_j 表示被试 j 的能力参数, β_i 表示题目 i 的难度参数。该模型将快速猜测作答的答对概率限定为 1, 这也适用于对题目有预了解的作答。如果将该模型用于不努力作答的情境, 根据其正确率低的特点, 可以将答对概率设为一个较低的值(如对于多项选择题, 设为随机水平)。

$\pi_{ij|z}$ 可以使用反应时来预测, 即

$$\pi_{ij|z}=\ln\frac{P\left(C_{ij}=1|time_{ij}\right)}{1-P\left(C_{ij}=1|time_{ij}\right)}=a+b\cdot time_{ij} \quad (2)$$

其中, a 和 b 表示预测被试 j 在题目 i 上的作答分组的截距和斜率。

该模型可以应用 Mplus 软件, 采用稳健标准误的极大似然估计方法估计参数(Pokropek, 2016)。Pokropek (2016)使用模拟研究证明该方法能够得到较准确的识别结果和参数估计结果。

2.1.2 半参数化的混合模型

Molenaar 等人(2018)提出了半参数化混合模型来区分快速作答和慢速作答。如果分类结果显示快速作答的反应时小于正常被试读题、理解和作答所需要的时间, 则可以认为所识别出的快速作答即为快速异常作答, 而慢速作答为正常作答。该方法假设在每个类别内部, 反应时服从对数正态分布。使用 p = 1,…, N 代表被试, i = 1,…, I 代表题目。C_{pi} 表示被试 p 在题目 i 上的作答类别, 假设 $C_{pi}=0$ 表示慢速作答, $C_{pi}=1$ 表示快速作答。被试 p 在 I 道题目上的分类为向量 $\boldsymbol{C}_p$ = $\left[C_{p1},C_{p2},\ldots,C_{pI}\right]$。观察到作答向量为 $\boldsymbol{x}_p$ =

$\left[X_{p1}, X_{p2}, \ldots, X_{pI}\right]$的概率为

$$P\left(\boldsymbol{x}_p | \theta_p, \boldsymbol{c}_p\right)=\prod_{i=1}^{I} \omega\left(\Omega_{pi}\right)^{X_{pi}} \omega\left(-\Omega_{pi}\right)^{1-X_{pi}} \quad (3)$$

其中

$$\Omega_{pi}=\left[\alpha_{0i}\left(1-C_{pi}\right)+\alpha_{1i} C_{pi}\right] \theta_p+\beta_{0i}\left(1-C_{pi}\right)+\beta_{1i} C_{pi} \quad (4)$$

θ_p是被试 p 的能力参数，ω(·) 是 logistic 方程，α_{si}是题目 i 在类别 s 的区分度参数(s=0,1)，β_{si}是题目 i 在类别 s 的容易度参数。

假设被试 p 在题目 i 上的连续反应时 T_{pi} 能够通过一定的转换关系得到类别变量$\dot{T}_{pi}$，即:

$$\dot{T}_{pi}=z \text{ 如果} k\left(T_{pi}\right) \in\left(b_{zi}, b_{(z+1)i}\right), z=0,1, \cdots, Z-1 \quad (5)$$

其中，b_{zi} 表示反应时转换的阈值，Z 表示反应时转换后的类别数，$k(\cdot)$ 表示转换函数。如果用虚无变量 d_{piz} 表示 $\dot{T}_{pi}$ 是否属于类别 z (d_{piz}=1 或者 d_{piz}=0)，可以使用广义线性 IRT 模型表示分类关系

$$b\left[E\left(d_{piz} | \tau_p, \boldsymbol{c}_p\right)\right]=\gamma_{zi}-\delta C_{pi}-\varphi_i \tau_p, \delta>0 \quad (6)$$

γ_{zi}表示题目 i 的反应时属于类别 z 的反应时类别参数，φ_i是斜率，τ_p是被试 p 的速度参数，δ是作答分类的系数。限定$\delta>0$是为了确保作答类别为$C_{pi}=1$的反应时分到低的反应时类别 z 中可能性更大，即反应时更短，因此$C_{pi}=1$表示快速作答，$C_{pi}=0$表示慢速作答。他们提出了两种链接函数 $b(\cdot)$，累积类别函数和相邻类别函数，用于预测反应时属于某个类别的概率。其中，累积类别函数类似于等级评分模型(Samejima, 1969)，相邻类别函数类似于分部计分模型(Masters, 1982)。例如，使用相邻类别函数，有

$$P\left(\dot{\boldsymbol{t}}_p | \tau_p, \boldsymbol{C}_p\right)=\prod_{i=1}^{I} \frac{\exp \left(\sum_{z=0}^{\dot{T}_{pi}} \gamma_{zi}-\delta C_{pi}-\varphi_i \tau_p\right)}{\sum_{j=0}^{Z-1} \exp \left(\sum_{z=0}^{j} \gamma_{zi}-\delta C_{pi}-\varphi_i \tau_p\right)} \quad (7)$$

其中类别参数γ_{zi}可以根据下式的限定得到

$$\sum_{z=0}^{Z-1}-\delta-\varphi_i \tau_p+\gamma_{zi}=0 \quad (8)$$

研究证明，当反应时转换后的类别数设定为 7、5 或 3 时，该方法能得到无偏的参数估计结果，相比于将反应时当作连续变量的方法，检验力几乎不受影响(Molenaar et al., 2018)。

2.1.3 基于反应时的混合作答反应模型

为了弥补半参数化的混合模型将反应时转换为分类变量的缺陷，Molenaar 和 de Boeck (2018)提出了基于反应时的混合作答反应模型以区分快速作答和慢速作答。

在反应时部分，参考 van der Linden (2006)的模型。使用 p = 1,…, N 代表被试，i = 1,…, I 代表题目，对于原始反应时 T_{pi}，有

$$\ln\left(T_{pi}\right)=\lambda_i-\tau_p+\varepsilon_{pi} \quad (9)$$

其中，λ_i表示题目 i 的时间密度参数，τ_p表示被试 p 的速度参数，ε_{pi}是残差项。

在作答反应部分，分别对快速作答和慢速作答定义不同的测量模型(Partchev & de Boeck, 2012)。即

$$\begin{aligned} & P\left(X_{pi}=1 | \theta_p, \alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i}\right)= \\ & \pi_{pi} P\left(X_{pi}=1 | \theta_p, \alpha_{0i}, \beta_{0i}\right)+ \\ & \left(1-\pi_{pi}\right) P\left(X_{pi}=1 | \theta_p, \alpha_{1i}, \beta_{1i}\right) \end{aligned} \quad (10)$$

其中，π_{pi}表示被试 p 在题目 i 上的作答属于类别 0 的概率，$1-\pi_{pi}$则表示被试 p 在题目 i 上的作答属于类别 1 的概率。α_{0i}, β_{0i} 和 α_{1i}, β_{1i} 分别表示类别 0 和类别 1 的作答在题目 i 上的区分度参数、难度参数。与两参数 IRT 模型一致，类别 0 和类别 1 的测量模型可以分别表示为

$$\operatorname{logit}\left[\left(X_{pi}=1 | \theta_p, \alpha_{0i}, \beta_{0i}\right)\right]=\alpha_{0i} \theta_p-\beta_{0i} \quad (11)$$

$$\operatorname{logit}\left[\left(X_{pi}=1 | \theta_p, \alpha_{1i}, \beta_{1i}\right)\right]=\alpha_{1i} \theta_p-\beta_{1i} \quad (12)$$

然后使用反应时来预测类别。被试 p 在题目 i 上的作答属于类别 0(C_{pi}=0)概率的 logit 为

$$\begin{aligned} & \operatorname{logit}\left[P\left(C_{pi}=0 | T_{pi}, \lambda_i, \tau_p, \sigma_{\varepsilon i}, \zeta_1, \zeta_0\right)\right]= \\ & \zeta_1\left(\frac{\ln\left(T_{pi}\right)-\left(\lambda_i-\tau_p\right)}{\sigma_{\varepsilon i}}-\zeta_0\right) \end{aligned} \quad (13)$$

其中，斜率参数$\zeta_1 \in[0, \infty)$以避免标签转移(指两个类别意义的转移)。被试 p 在题目 i 上的实际反应时与模型预测均值相比越长，越可能被分到类别 0 中。因此，类别 0 代表慢速作答，类别 1 表示快速作答。截距参数ζ_0表示作答被分到慢速作答类别的难度参数。模拟研究证明，应用贝叶斯框架下的 MCMC 算法，该模型能够得到较准确的参数估计结果(Molenaar & de Boeck, 2018)。

2.2 基于反应时和作答反应的混合多层模型

2.2.1 混合多层模型

van der Linden (2007)的多层模型是迄今最流

行的, 基于作答反应和反应时的多层模型。该模型包括两个水平, 第一水平是测量模型, 包括作答反应部分的 IRT 模型和反应时部分的标准对数正态分布模型。第二水平是个体水平, 通过能力和速度的协方差结构, 将作答反应和反应时联系起来。

具体来看, 第一水平的模型可以表示为

$$\begin{cases} P\left(Y_{ij}=1\middle|\theta_i\right)=\dfrac{\exp\left(a_j\left(\theta_i-b_j\right)\right)}{1+\exp\left(a_j\left(\theta_i-b_j\right)\right)} & \text{作答反应模型} \\ \ln\left(t_{ij}\right)\middle|\tau_i\sim N\left(\beta_j-\tau_i,\alpha_j^{-2}\right) & \text{反应时模型} \end{cases} \quad (14)$$

其中, $P\left(Y_{ij}=1\middle|\theta_i\right)$表示被试 $i(i = 1,\ldots, I)$在题目 j $(j = 1,\ldots, J)$上正确作答的概率, t_{ij} 表示被试 i 在题目 j 上的反应时, a_j 和 b_j 分别是题目 j 的区分度参数和难度参数, β_j 表示题目 j 的时间密度参数, α_j 表示题目 j 的时间区分度参数。时间密度类似于 IRT 中难度的概念, 时间密度越大, 完成题目所需要的时间就越长, 而时间区分度类似于 IRT 中区分度的概念, 时间区分度越大, 不同速度被试在题目上期望反应时的差异就越大。$N()$表示正态分布, θ_i 和 τ_i 是被试 i 的能力参数和速度参数。在第二水平(个体水平), 假设被试参数 $\xi_i=\left(\theta_i,\tau_i\right)$ 服从二元正态分布 $\xi_i\sim N\left(\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p\right)$, 其中均值向量为 $\boldsymbol{\mu}_p=\left(\mu_\theta,\mu_\tau\right)$, 协方差矩阵为

$$\boldsymbol{\Sigma}_p=\begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau}\\ \sigma_{\tau\theta} & \sigma_\tau^2\end{pmatrix} \quad (15)$$

为了模型识别, 对于作答反应模型通常限定 $\mu_\theta = 0, \sigma_\theta^2=1$。对于反应时模型, 可以限定速度参数的均值或者时间密度参数的均值。Wang 和 Xu (2015)建议限定 $\mu_\tau=0$ 以便于和作答反应模型的限定保持 一致。

这一模型的优势为在同一模型中协调了速度和能力的关系, 因此, 反应时信息可以帮助提高作答反应模型参数估计准确性, 反过来, 作答反应信息也可以帮助提高反应时模型参数估计准确性(van der Linden, 2007)。

在此基础上, Wang 和 Xu (2015)提出了基于反应时和作答反应的混合多层模型(mixture hierarchical model), 用于识别异常作答。根据正常作答行为和异常作答行为的特点, 可以对总体的作答反应模型和反应时模型进行分解。

在作答反应模型部分, 被试 i 在题目 j 上答对的概率为

$$P\left(Y_{ij}=1|\Delta_{ij}\right)=$$
$$\left(1-\Delta_{ij}\right)P\left(Y_{ij}=1|\Delta_{ij}=0\right)+\Delta_{ij}P\left(Y_{ij}=1|\Delta_{ij}=1\right) \quad (16)$$

其中, Δ_{ij} 是表示作答行为分类的潜变量, $\Delta_{ij}=1$, 表示被试 i 回答题目 j 是异常作答, $\Delta_{ij}=0$, 表示是正常作答。如果 $\Delta_{ij}=0$, 可使用三参数 logistic (3PL)模型预测正常作答的答对概率。

$$P\left(Y_{ij}=1|\Delta_{ij}=0,a_j,b_j,c_j\right)=$$
$$c_j+\left(1-c_j\right)\frac{\exp\left[a_j\left(\theta_i-b_j\right)\right]}{1+\exp\left[a_j\left(\theta_i-b_j\right)\right]} \quad (17)$$

其中, a_j,b_j 和 c_j 分别代表题目 j 的区分度参数、难度参数和猜测参数。θ_i 表示被试 i 的能力参数。

根据测验性质和作答类型的不同, 其他的 IRT 模型, 例如两参数 logistic (2PL)模型、分部计分模型或者等级评分模型也可以应用于这一混合多层模型的框架下。如果 $\Delta_{ij}=1$, 被试 i 回答题目 j 是异常作答, 答对概率是 g_j。即

$$P\left(Y_{ij}=1|\Delta_{ij}=1\right)=g_j \quad (18)$$

注意这里的 g_j 与三参数 logistic (3PL)模型中猜测参数 c_j 的含义不同。g_j 表示被试异常作答的正确率。而 c_j 反映了被试正常作答条件下的猜测正确率。

在反应时模型部分, 假设对于被试 i 和题目 j, 观察到的反应时 T_{ij}^{obs} 可以表示为

$$T_{ij}^{obs}=\left(1-\Delta_{ij}\right)T_{ij}+\Delta_{ij}C_{ij} \quad (19)$$

其中, T_{ij} 表示被试 i 正常作答题目 j 所需的时间, C_{ij} 表示被试 i 异常作答题目 j 所需的时间。

假定正常作答行为的反应时服从对数正态分布(van der Linden, 2007)。

$$\ln\left(T_{ij}\right)\sim N\left(\beta_j-\tau_i,\left(\frac{1}{\alpha_j}\right)^2\right) \quad (20)$$

其中, β_j 是题目 j 的时间密度参数, α_j 是题目 j 的时间区分度参数, τ_i 是被试 i 的速度参数。

假定异常作答行为的反应时也服从对数正态分布

$$\ln\left(C_{ij}\right)\sim N\left(\mu_c,\sigma_c^2\right) \quad (21)$$

这个分布的均值(μ_c)和方差(σ_c^2)对于所有的

被试和题目都相同，用于反映异常作答提供的心理测量学信息少的特点。

与 van der Linden (2007)的多层模型一致，该混合模型包含三个局部独立性假设。第一，基于被试的能力水平和是否正常作答的分类，作答反应具备条件独立性。第二，基于被试的速度水平和是否正常作答的分类，反应时具备条件独立性。第三，基于被试参数(能力参数、速度参数)和是否正常作答的分类，对于每道题目来说，作答反应和反应时具备条件独立性。

Wang 和 Xu (2015)采用基于蒙特卡洛的 EM 算法(Monte Carlo-based EM algorithm, MCEM)估计参数。这一算法是在标准 EM 算法的基础上，通过蒙特卡洛模拟的方式得到 E 步骤的期望值。在 MCEM 的每次迭代中，取得一个蒙特卡洛样本最方便的方式就是使用MCMC算法，通常包括Gibbs 抽样或者 MH (Metropolis–Hastings, MH)抽样。后来，Wang 等人(Wang, Xu, & Shang, 2018; Wang, Xu, Shang, & Kuncel, 2018)又直接采用了贝叶斯框架下的 MCMC 算法得到参数的后验分布，进而计算后验均值得到参数的点估计值。后面 2.2.2—2.2.4 中介绍的模型都采用该方法实现参数估计。这类估计方法的优势主要有两个方面。一是它允许针对不同类型的异常作答，对模型中的参数加入特定的先验分布，以限定参数估计值的大致范围。例如，我们可以限定快速异常作答反应时的均值 μ_c 为一个均值相对较小的分布，用以表示其反应时短的特点。又例如，快速猜测作答和加速作答的 g_j 应当限定为小于正常作答使用 3PL 模型得到的答对概率值，而对题目有预了解的作答的 g_j 应当限定为大于正常作答使用 3PL 模型得到的答对概率值。二是对于每个参数可以得到其后验分布，便于基于整个后验分布而不是点估计值进行后续的统计检验(如后验预测 p 值，posterior predictive p-value, PPP 等)。

Wang 和 Xu (2015)的模拟研究结果证明，当数据中同时含有正常作答与异常作答时，应用混合多层模型相比于传统多层模型能够得到更准确的参数估计结果。Wang, Xu, Shang 和 Kuncel (2018)的研究证明，无论数据是基于混合多层模型还是残差模型产生，混合多层模型在正确识别率和错误拒绝率上表现都较好，特别是当异常作答的比例较高时，该模型相比于贝叶斯残差法优势更加明显。

2.2.2 应用于高阶 IRT 的混合多层模型

Lu 等人(2020)近期又将混合多层模型拓展应用于高阶 IRT 模型，主要处理测验结构为题目间多维的情况。这一模型在作答反应模型部分采用高阶 IRT 模型，即对于被试 i 在分维度 v(v=1,2,3,...V, 共 V 个分维度)上的能力 $\theta_{iv}^{(1)}$，有如下线性关系

$$\theta_{iv}^{(1)} = \beta_v \cdot \theta_i^{(2)} + \varepsilon_{iv}^{(1)} \tag{22}$$

其中，$\theta_i^{(2)}$ 表示被试 i 的高阶能力，β_v 表示 $\theta_{iv}^{(1)}$ 的回归系数，$\varepsilon_{iv}^{(1)}$ 表示 $\theta_{iv}^{(1)}$ 的残差项。基于模型识别的考虑，假设 $\theta_i^{(2)} \sim N(0,1)$，并且 $\varepsilon_{iv}^{(1)} \sim N\left(0, 1-\beta_v^2\right)$。这样的限定能够保证高阶能力和低阶能力在同一尺度上。η_{ijv} 为表示作答是否为正常作答的指标变量，其值为 1 表示异常作答，0 表示正常作答。当被试 i 在分维度 v 的题目 j 上的作答为正常作答时($\eta_{ijv} = 0$)，可以使用三参数正态肩型模型(也可以使用其他 IRT 模型)拟合数据，即

$$P\left(Y_{ijv} = 1 | \eta_{ijv} = 0, \theta_{iv}^{(1)}, a_{jv}, b_{jv}, c_{jv}\right) = c_{jv} + \left(1 - c_{jv}\right) \cdot \Phi\left(a_{jv} \cdot \left(\theta_{iv}^{(1)} - b_{jv}\right)\right) \tag{23}$$

其中，$\Phi(\cdot)$ 表示标准正态分布函数，a_{jv}, b_{jv} 和 c_{jv} 分别表示分维度 v 上题目 j 的区分度参数，难度参数和猜测参数。

该模型关于异常作答的正确作答概率，以及反应时部分模型的分解，均与 Wang 和 Xu (2015)的混合多层模型一致。Lu 等人(2020)在多维测验，且每个维度的题目数，时间限制不同的情境下模拟数据，比较了这一模型和基于单维 IRT 模型的混合多层模型的表现。结果证明，该模型的参数估计偏差更小，识别准确性更高。

2.2.3 基于混合多层模型的两步方法

针对同时存在不努力作答和对题目有预了解的情境，Wang, Xu 和 Shang (2018)在混合多层模型的基础上，又提出了确定异常行为模式的两步方法。

具体来说，第一步是将数据与混合多层模型拟合。第二步是通过对作答模式的检验，进一步确定异常作答是不努力作答还是对题目有预了解的作答。

第二步的检验方法为，首先，对于异常作答的题目进行汇总，即对于被试 i，计算异常作答

($\Delta_{ij}=1$)的题目数 J_i。然后, 计算每名被试标准化残差的均值(Wright & Stone, 1979)。

$$V_i(\theta)=\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{Y_{ij}-P_j\left(\hat{\theta}_i\right)}{\sqrt{P_j\left(\hat{\theta}_i\right)\left[1-P_j\left(\hat{\theta}_i\right)\right]}} \tag{24}$$

其中, $P_j\left(\hat{\theta}_i\right)$是基于被试 i 的能力参数估计值 $\hat{\theta}_i$, 代入 IRT 模型计算得到的被试 i 在题目 j 上正常作答的期望概率。由于这个式子中 $\hat{\theta}_i$ 的点估计值可能不准确, 因此 Wang, Xu 和 Shang (2018)使用贝叶斯方法来改进。即, 使用 $P\left(1|y_i^*\right)$ 替代 $P_j\left(\hat{\theta}_i\right)$

$$P\left(1|y_i^*\right)=\frac{1}{H_{ij}}\int P_j(\theta)\prod_{k\in R_{-J_i}}P_k(\theta)^{y_{ik}}\left[1-P_k(\theta)\right]^{1-y_{ik}}g(\theta)d\theta \tag{25}$$

其中, y_i^* 表示被试 i 在正常题目($\Delta_{ij}=0$)上的作答反应, $P_k(\theta)$是基于 3PL 模型计算的被试 i 在第 k 道正常作答的题目上的正确率, y_{ik} 表示被试 i 在第 k 道正常作答的题目上的实际作答, R_{-J_i} 表示被试 i 正常作答的题目。$g(\theta)$表示 θ 的先验密度, H_{ij} 是被试 i 在题目 j 上的正态化常数。

最后确定阈值 v, 如果 $V_i(\theta)>v$, 被试 i 的异常作答是对题目有预了解, 如果 $V_i(\theta)<-v$, 被试 i 的异常作答是不努力作答, 如果 $-v<V_i(\theta)<v$, 被试 i 的异常作答混合了以上两种模式。他们的模拟研究(Wang, Xu, & Shang, 2018)探讨了阈值 v 的选取问题, 建议在实践中选择 v=0。研究证明, 基于混合多层模型的两步方法不仅能够在各种条件下得到较高的正确识别率和较低的错误拒绝率, 还能够得到较准确的参数估计结果。

2.2.4　考虑了缺失数据的混合多层模型

针对同时存在不努力作答和缺失的情境, 基于混合多层模型, Ulitzsch 等人(2020)提出了考虑了缺失数据的混合多层模型。这一模型的基本框架是将作答先分为正常作答和不努力作答, 其中不努力作答又有忽略题目和随机猜测作答两种表现。

他们的模型中加入了潜变量 ϕ_i 用以表示被试 i 的努力程度。使用 Rasch 模型来预测被试是否努力作答的概率, 可以得到

$$P\left(\Delta_{ij}=1\right)=\frac{\exp\left(\phi_i-\iota_j\right)}{1+\exp\left(\phi_i-\iota_j\right)} \tag{26}$$

其中, ι_j 表示题目 j 的努力程度难度, 类似于 IRT 中对难度的定义, ι_j 越高, 表示被试在这道题目上越不容易努力作答, Δ_{ij} 为是否努力作答的二分变量($\Delta_{ij}=1$ 表示努力作答, $\Delta_{ij}=0$ 表示不努力作答)。他们还定义了一个表示作答是否缺失的二分变量 d_{ij}, $d_{ij}=1$ 表示被试 i 在题目 j 上无作答, $d_{ij}=0$ 表示被试 i 在题目 j 上有作答。如果被试 i 在题目 j 上是努力作答($\Delta_{ij}=1$), 则 $P\left(d_{ij}=1|\Delta_{ij}=1\right)=0$, $P\left(d_{ij}=0|\Delta_{ij}=1\right)=1$, 即被试 i 在题目 j 上肯定有作答。此时可参考 van der Linden(2007)的多层模型拟合作答反应和反应时。如果被试 i 在题目 j 上是不努力作答($\Delta_{ij}=0$), 那么 $d_{ij}=1$ 表示被试 i 在题目 j 上是由于忽略而缺失, $d_{ij}=0$ 表示被试 i 在题目 j 上是随机猜测。则有

$$P\left(d_{ij}=1|\Delta_{ij}=0\right)=\frac{\exp\left(\gamma_0+\gamma_1\theta_i+\gamma_2\tau_i\right)}{1+\exp\left(\gamma_0+\gamma_1\theta_i+\gamma_2\tau_i\right)} \tag{27}$$

其中, θ_i 和 τ_i 分别表示被试 i 的能力参数和速度参数, γ_0 和 γ_1,γ_2 分别是截距和斜率参数。对于随机猜测作答, 答对的概率为

$$P\left(u_{ij}=1|d_{ij}=0,\Delta_{ij}=0\right)=c \tag{28}$$

其中, c 是猜测参数。

在反应时部分, 与 Wang 和 Xu(2015)的模型一致, 不努力作答的反应时服从均值(β_D)和方差(σ_D^2)恒定的对数正态分布, 即

$$ln\left(t_{ij}|\Delta_{ij}=0\right)\sim N\left(\beta_D,\sigma_D^2\right) \tag{29}$$

此外, 还假定不努力作答所需要的反应时小于努力作答, 因此, 努力作答题目的时间密度参数(β_j)和不努力作答的对数反应时均值(β_D)存在以下关系

$$\beta_j=\beta_D+\beta_j^*, \text{其中} \beta_j^*\geqslant 0 \tag{30}$$

其中, β_j^* 表示对于题目 j, 努力与不努力作答相比多花的时间。

最后, 假设所有被试参数服从均值为

$$\boldsymbol{\mu}_{\mathrm{P}}=\left(\mu_\phi,\mu_\theta,\mu_\tau\right) \tag{31}$$

协方差矩阵为

$$\boldsymbol{\Sigma}_{\mathrm{P}}=\begin{pmatrix}\sigma_\phi^2 & \sigma_{\phi\theta} & \sigma_{\phi\tau}\\ \sigma_{\theta\phi} & \sigma_\theta^2 & \sigma_{\theta\tau}\\ \sigma_{\tau\phi} & \sigma_{\tau\theta} & \sigma_\tau^2\end{pmatrix} \tag{32}$$

的多元正态分布。

综上, 该模型框架可以用图 1 表示。

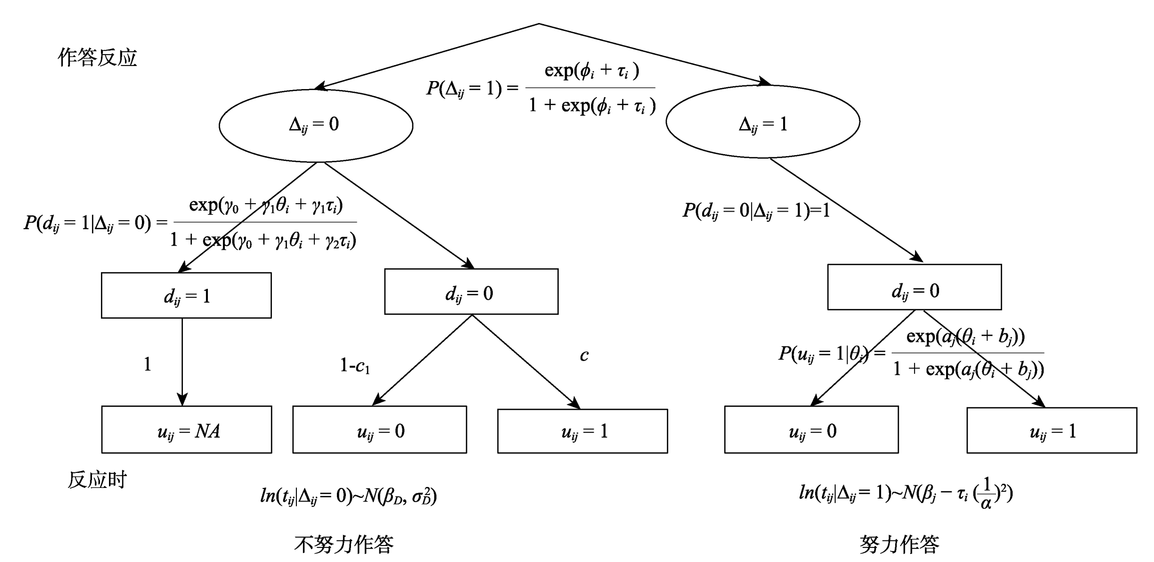


图 1 考虑了缺失的混合模型示意图

模拟研究证明，在不努力作答比例较高的条件下，该模型参数估计的返真性较好(Ulitzsch et al., 2020)。

2.3 混合模型法简评

混合模型在处理异常作答时最大的优势在于能够同时完成异常作答的识别和模型参数估计。具体来看，各方法具有如下局限性。

首先，等级分组的反应时模型方法包含强假设，即对于所有题目所有被试，快速猜测作答的答对概率为 1。这显然不尽合理。Wang 和 Xu (2015)的混合多层模型就弱化了这一假设，限定每道题目异常作答的答对概率为 g_j。其次，半参数化的混合模型和基于反应时的混合作答反应模型都用于区分快速作答和慢速作答，其识别快速异常作答的有效性尚待验证。此外，对于不同类别，这些模型需要估计的参数数量是相同的。因此如果将它们用于识别异常作答，可能由于异常作答的比例明显过小，造成该类别参数估计的标准误偏大，而影响其检验力(Molenaar et al., 2018)。Wang 和 Xu (2015)的混合多层模型以及在此基础上拓展的应用于高阶 IRT 的多层混合模型、基于混合多层模型的两步方法和考虑了缺失数据的混合多层模型，对异常作答的作答反应和反应时模型设置了相对较少的待估参数(例如 g_j, μ_c, σ_c^2)，能够在一定程度上解决这一问题。然而该类模型面临的主要质疑仍然是混合模型普遍存在的强假设问题。虽然与等级分组的反应时模型直接限定快速猜测作答的答对概率相比，这类模型的假设有一定弱化，但是仍有不尽合理之处。例如，这类模型假设异常作答行为的反应时服从均值和标准差恒定的对数正态分布。然而实际中异常作答可能和被试因素(例如，学业能力，作答速度等)，或者题目因素(例如题目位置，题型等)相关(Goldhammer et al., 2016; Lee & Jia, 2014; Molenaar, Bolsinova et al, 2016; Molenaar et al., 2018; Molenaar, Oberski et al., 2016)。因此，对于异常作答反应时分布的假设可能不合理。又例如，这类模型假设异常作答的正确率为 g_j，即对于不同被试，在同样题目上，异常作答行为的答对概率是相同的。但是 Feinberg 和 Jurich (2018)发现，不同能力水平被试异常作答的正确率不同。因此这一假设不尽合理。综上，当混合多层模型的假设无法满足时，这种方法可能无法成功划分作答行为的不同类型(Molenaar et al., 2018; Ranger & Kuhn, 2017)。除此之外，考虑了缺失数据的混合多层模型非常复杂，待估参数偏多，存在模型拟合时间长(例如，对于 1000 人在 20 道题上的作答，模型拟合所需时间在 24 小时以上，Ulitzsch et al., 2020)，参数估计不易收敛等问题。

3 三类方法的综合分析与比较

3.1 三类方法基本思路的分析与比较

从功能来说，反应时阈值法和反应时残差法都是识别快速异常作答，之后需要采用降低权重的方式进行参数估计。而混合模型法在建模时就

考虑了不同作答特点, 能够一次性解决异常作答的识别和模型参数估计的问题。

从思路来说, 尽管三类方法都假设, 如果存在异常作答, 整个作答反应和反应时都呈现出混合两类模式的特点。但是, 它们处理两类作答模式的思路是不同的。反应时阈值法和反应时残差法首先关注所有作答的反应时分布。再根据快速异常作答具有反应时短的特点, 从整体分布中划分阈值, 将阈值之外的异常值识别为快速异常作答。这类似于假设检验的思路。在整个分布中极端的数值, 仍属于这个分布的概率是非常小的。因此我们有更充分的理由相信这些极端的数值是属于另一个分布的(快速异常作答的反应时分布)。从这个意义上说, 反应时残差法也是通过在反应时残差或期望反应时分布上设定阈值来做出判断的。因此, 在更广泛的层面上, 反应时残差法也可以看作一种"反应时阈值法"。然而, 当整个数据中已经混有快速异常作答时, 这些异常的作答模式必然会影响整个数据反应时的分布。例如, 这种情况下标准化反应时残差其实可能已经不符合标准正态分布了。采用标准化反应时残差法本身的前提假设就不成立, 会造成识别结果的偏差。而混合模型法的基本思想在于用平等的视角对待两类作答模式。将作答反应的正确率, 反应时分布, 都视作两类作答模式的混合。这种思路具有一定的灵活性。一是在数据中存在异常作答的情况下, 两类作答分别对各自的模型参数估计提供信息, 不会出现像传统模型那样, 随着异常作答比例增加, 模型参数估计误差增大的现象。二是在数据中不存在异常作答的情况下, 相当于每个作答的潜类别都相同, 从理论上说该模型也能够处理这种数据。

3.2　三类方法局限性的分析与比较

总的来说, 三类方法各有优缺点。反应时阈值法原理简单, 便于应用和操作, 是最早提出的一类方法。但是这类方法由于缺乏背后理论模型的支持, 在现实中无法确定阈值的情况屡有发生, 因此其科学性也受到越来越多的质疑。反应时残差法基于反应时模型构造出期望分布, 具有一定的理论依据。但是当存在快速异常作答时, 传统的反应时模型是否拟合良好, 计算得到期望分布是否符合理论分布, 也是值得反复斟酌的问题。混合模型法也基于特定的理论模型, 同时考虑了不同类型数据的特点, 在一定程度上突破了前两种方法的局限性。并且, 该方法可以通过灵活设定异常作答部分参数的先验分布, 应用于不同类型异常作答的识别。然而, 混合模型的方法还普遍存在包含强假设, 计算复杂耗时长, 有时参数估计不收敛等缺陷。表 1 总结了本文中介绍的方法的主要局限性。

表 1　本文中所有方法的主要局限性总结

方法类型	具体方法	没有综合利用反应时和作答反应的信息	没有基于理论分布	偶有例外, 无法批量应用	包含有关异常作答的强假设	对高比例异常作答敏感	异常作答比例低时容易出现问题	计算复杂耗时长	识别结果不一定是异常作答	只能用于已知异常作答对概率的情境	只能用于识别快速异常作答
反应时阈值法	统一阈值法	×	×								×
	根据题目特征求阈值法	×	×								×
	双峰分布交点求阈值法	×	×	×							×
	常模阈值法		×								×
	基于信息求阈值法		×	×							×
	条件分布法		×	×						×	×
反应时残差法	标准化反应时残差法	×				×					×
	贝叶斯残差法					×		×			×
混合模型法	等级分组的反应时模型				×					×	×
	半参数化的混合模型				×		×	×	×		×
	基于反应时的混合作答反应模型				×		×	×	×		×
	基于反应时和作答反应的混合多层模型				×		×	×			

注: 表中的×表示方法有此项局限性。

总的来说，混合模型法的局限性主要来自于三个方面。一是包含一些关于异常作答的作答正确率和反应时分布的强假设，如果这些假设遭到违背，可能无法得到准确的识别结果。二是异常作答比例较低时容易出现问题。例如，当异常作答的比例较小或者样本量较小时，有时很难得到收敛的结果(Ranger et al., 2019)。当数据中不存在异常作答时，甚至会出现模型识别的问题(Dolan et al., 2002)。三是计算复杂耗时长。例如，即便使用贝叶斯框架下的 MCMC 算法估计参数的后验分布，在先验分布设置合理的情况下，也需要较长时间。笔者借助普通计算机(处理器为 i7-4500U 内存为 8GB)，对样本量为 2000，题目数为 30，异常作答比例约为 25%的模拟数据，基于 Wang 和 Xu (2015)的混合模型，应用贝叶斯框架下基于 Gibbs 抽样的 MCMC 算法估计参数后验分布，迭代收敛所需时间达到 9 小时以上。

由于不同类型的方法具有不同特点，因此在实际的心理与教育测验数据分析中，应当结合具体测验的特点以及要处理的异常作答类型，选用合适的方法。例如，在一些高利害的测验中，学生往往具有较高的动机，考试安全性也较高，异常作答的现象很少，并且主要表现为快速猜测等快速异常作答。这时低比例的快速异常作答对传统模型参数估计结果的影响很小，可以选用反应时残差法，或反应时阈值法识别快速异常作答并在估计时降低权重。而在一些低利害的测验中，异常作答发生的频率较高，并且主要表现为不努力作答。这时反应时残差法会出现较大偏差，可以选用对高比例异常作答不敏感的混合模型法，一次性解决识别和参数估计的问题。

4 问题与展望

目前，几乎所有的心理与教育测量模型都建立在学生正常作答的前提假设下(Wise, 2015)，并没有考虑异常作答可能对个人分数等造成影响。有很多研究者提出，如果能够建立一套程序证明个人分数的效度(ISV, individual score validity)，就有责任在分析数据之前使用这套程序(Hauser & Kingsbury, 2009; Hauser et al., 2008)。处理异常作答显然就是这套程序的一部分。

混合模型虽然在心理与教育测量中早有应用，但是在很长一段时期内，都仅停留在个人层面的分类。随着对数据分析精度要求的提高，以及对数据中有效信息充分利用的需求不断增加，实现作答层面的分类成为了混合模型发展的重要方向之一。关于结合了反应时与作答反应的多层模型的深入研究和推广应用，又为综合利用多元信息识别和分析异常作答提供了重要的模型基础。而贝叶斯框架下的 MCMC 算法在心理与教育测量中的广泛应用，又使得更为复杂的混合模型的参数估计得以顺利实现。可以说，混合模型法的出现，是模型和估计方法发展优化的共同结果。虽然该方法在异常作答的处理中具有种种优势，但它毕竟是一类较新的方法，本身也具有一定的局限性。因此无论是方法改进、方法适用性的理论研究，还是方法在实际中应用的实践研究，都有着较为广阔的发展空间。以下对混合模型方法未来可能的研究方向提供一些建议，供感兴趣的研究者参考。

4.1 检验违背前提假设时模型的稳健性

众所周知，混合模型最为研究者诟病的方面是它含有一些强假设。正是由于强假设的存在，才使得对分类潜变量、不同类别模型参数的估计成为可能。而另一方面，这些强假设也在一定程度上增加了模型在假设不满足时估计结果不理想的风险。Wang, Xu, Shang 和 Kuncel (2018)曾在数据基于混合模型假设产生和基于残差模型产生的条件下，对混合多层模型和贝叶斯残差法进行比较。研究结果在一定程度上证明了不管基于何种模型产生数据，混合多层模型相比于贝叶斯残差法在异常作答的识别和参数估计结果返真性上都表现出较大的优势。但是，混合多层模型在拟合基于残差法产生的数据时的表现，要差于基于混合多层模型产生的数据。然而，在他们的模拟研究中，基于残差法产生异常作答的反应时数据仅违背了混合多层模型中关于反应时模型的假设，异常作答的答对概率仍符合其假设。除此之外，混合多层模型还包含三个局部独立性假设(见本文 2.2.1)，在已有的混合多层模型研究中，这些假设都是满足的。今后应针对混合模型各种前提假设遭到违背的情况开展广泛的模拟研究，探讨该方法的稳健性。

4.2 固定部分题目参数以提高方法估计速度

即使应用了贝叶斯框架下的 MCMC 算法，一些较为复杂的混合模型仍面临着计算复杂耗时长

的问题。这是因为在迭代过程中, 所有参数都需要从后验分布中抽取。可以设想, 如果已知部分参数(如题目参数), 将其固定对其余参数进行条件估计, 应当能够有效提高估计速度。为了得到准确的题目参数估计结果, 可以应用 Liu 等人(2020)提出的对被试个体分类的混合模型方法, 先筛选出正常作答的被试群体, 基于这一群体拟合 van der Linden (2007)的多层模型, 得到较准确的正常作答部分的题目参数估计结果。再将这些题目参数估计结果固定, 代入混合多层模型的参数估计过程, 可以明显缩短估计时间。经笔者实验证明, 对于样本量为 2000, 题目数为 30, 快速异常作答比例约为 25%的模拟数据, 基于混合多层模型(正常作答使用两参数 logistic 模型拟合), 采用贝叶斯框架下基于 Gibbs 抽样的 MCMC 算法估计参数后验分布, 应用这种固定部分题目参数估计的方式, 能够将估计时间缩短到原来的一半以下。

4.3　结合其他反应时模型以提高方法适用性

目前用于处理异常作答的混合模型在反应时部分多采用的是 van der Linden (2006, 2007)的反应时模型。尽管该模型可以算作应用最广泛的反应时模型, 但是, 也有很多研究者提出了一些其他的模型, 并认为这些模型在某些情况下具有更好的适用性。例如, 在实验心理学中较常用的三参数反应时模型(e.g., Cousineau, 2009), 反应时的半参数化模型(Wang, Chang et al., 2013; Wang, Fan et al., 2013), 在 van der Linden (2007)模型的基础上考虑了残差相关的模型(Bolsinova & Tijmstra, 2019)等。此外, Wang 和 Xu (2015)也指出, 目前的多层模型隐含了测验中只含有单一题型的假设。如果测验中含有多种题型, 时间密度参数可能依赖于具体的题型, 这可能需要在反应时模型部分允许不同题型的时间密度参数有不同的分布形态。因此, 如何基于其他的反应时模型构建相应的混合模型, 也是未来研究方向之一。

4.4　考虑实际复杂情境以提高方法针对性

目前的大多数研究都考察了仅存在一种类型异常作答的情境下, 混合模型法的有效性。然而在实际的心理与教育测验中, 往往不可能仅存在一种类型的异常作答。被试的复杂性常会带来数据情况的复杂性, 现实中测验所得到数据往往包含更复杂的问题, 也对识别异常值的统计方法提出了新的挑战。虽然已有一些研究者对这种复杂情境的处理开展了一些尝试。例如针对同时含有缺失数据和不努力作答的复杂情境, Ulitzsc 等人(2019)提出了考虑了缺失数据的混合多层模型。未来研究也可以拓展到数据同时包含忽略题目, 加速作答, 快速猜测作答, 对题目有预了解的作答等情况的复杂情境, 探索如何建立更具针对性的混合模型, 并考察如何解决这类模型的识别和参数估计(包括收敛等)等问题。

4.5　增加选择流程以提高方法使用效率

从已有研究结果来看, 混合模型法有一定的适用条件。当数据中异常作答的比例较高时, 使用该方法能够得到较准确的识别结果和参数估计结果, 方法使用效率高。而当数据中异常作答的比例较低时, 不仅会影响异常作答部分模型参数估计结果的准确性, 甚至还可能得到不收敛的结果(Ranger et al., 2019)。此时不仅方法使用效率低, 还可能根本不能应用, 可以改为选用其他对低比例异常作答不敏感的方法(如标准化反应时残差法)。然而在实际的数据清理中, 我们只能从测验是否为低利害测验, 测验的保密程度, 考生的基本情况, 以及监考反馈等方面, 大致推测异常作答的严重程度, 选择合适的方法。今后的研究可以尝试构建测量整个数据中异常作答严重程度的指标, 建立指标与使用混合模型法得到的参数估计结果准确性提高程度之间的联系。从而指导实践研究者根据指标反映出的数据污染情况选择合适的方法, 提高方法的使用效率。

参考文献

黄美薇, 潘逸沁, 骆方. (2020). 结合选择题与主观题信息的两阶段作弊甄别方法. *心理科学,* (1), 75–80.

简小珠, 焦璨, Steven P Reise, 彭春妹. (2010). 四参数模型对被试作答异常现象的拟合与纠正. *心理科学进展, 18*(3), 537–544.

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68*(1), 139–151.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self–report methodology. *Psychological Assessment*, *4*(3), 340–345.

Bolsinova, M., & Tijmstra, J. (2019). Modeling differences between response times of correct and incorrect responses. *Psychometrika*, *84*(4), 1018–1046.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Borghans, L., & Schils, T. (2012). *The leaning tower of PISA: Decomposing achievement test scores into cognitive and noncognitive components* (Unpublished doctorial dissertation). Maastricht University.

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41*(2), 137–148.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment, 15*(2), 223–234.

Cousineau, D. (2009). Fitting the three-parameter Weibull distribution: Review and evaluation of existing and new methods. *IEEE Transactions on Dielectrics and Electrical Insulation*, *16*(1), 281–288.

Custer, M., Sharairi, S., & Swift, D. (2012, April). *A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilizing the Rasch model and joint maximum likelihood estimation*. Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, BC, Canada.

Dolan, C. V., van der Maas, H. L. J., & Molenaar, P. C. M. (2002). A framework for ML estimation of parameters of (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods, Instruments & Computers*, *34*, 304–323.

Feinberg, R., & Jurich, D. (2018, April). *Using rapid responses to evaluate test speededness*. Paper presented at the meeting of the National Council of Measurement in Education (NCME), New York, NY.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers, No. 133). Paris, France: OECD Publishing.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183.

Hauser, C., & Kingsbury, G, G. (2009). *Individual score validity in a modest-stakes adaptive educational testing setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hauser, C., Kingsbury, G. G., & Wise, S. L. (2008). *Individual validity: Adding a missing link*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Hong, M. R., & Cheng, Y. (2019a). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods*, *51*(2), 573–588.

Hong, M. R., & Cheng, Y. (2019b). Clarifying the effect of test speededness. *Applied Psychological Measurement*, *43*(8), 611–623.

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive Assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397–419.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619.

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*(8), 1–24.

Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, *80*(4), 775–807.

Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 261–288.

Ma, L., Wise, S.L., Thum, Y.M., & Kingsbury, G. (2011, April). *Detecting response time threshold under the computer adaptive testing environment*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538.

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing*, *20*(3), 187–205.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205–228.

Molenaar, D., Bolsinova, M., Rozsa, S., & de Boeck, P. (2016).

Response mixture modeling of intraindividual differences in responses and response times to the Hungarian WISC-IV block design test. *Journal of Intelligence*, *4*(3), 10–29.

Molenaar, D., Oberski, D., Vermunt, J., & de Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626.

Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279–297.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods and Applications, 15*, 271–293.

Partchev, I., & de Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32.

Patton, J. M., Cheng, Y., Hong, M. R., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, *44*(3), 309–341.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS technical report for reading? Scaling results of starting cohort 3 in fifth grade*. NEPS Working Paper No. 15. Bamberg: Otto-Friedrich-Universitt, Nationales Bildungspanel.

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47.

Ranger, J., & Kuhn, J. T. (2017). Detecting unmotivated individuals with a new model-selection approach for Rasch models. *Psychological Test and Assessment Modeling*, *59*(3), 269–295.

Ranger, J., Wolgast, A., & Kuhn, J. T. (2019). Robust estimation of the hierarchical model for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 83–107.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?. *International Journal of Testing*, *17*(1), 74–104.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Unpublished doctorial dissertation). Friedrich-Schiller-University, Jena.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*(3), 795–819.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph Supplement No. 17). Richmond, VA: Psychometric Society.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232.

Schnipke, D. L. & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Mahwah, NJ: Lawrence Erlbaum.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34–49.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*(4), 1118–1141.

Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *TRAMES: A Journal of the Humanities & Social Sciences*, *17*(4), 433–448.

Sinharay, S., & Johnson, M. S. (2019). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 397–419.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 83–112.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 144–168.

Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for

response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223–254.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test–taking effort. *Applied Psychological Measurement, 41*(2), 115–129.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252.

Wise, S. L. (2017). Rapid - guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61.

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, *32*(4), 325–336.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86–105.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*. Chicago, IL: MESA Press.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology, 22*(1), 51–68.

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods, 24*(5), 658–674.

Mixture Model Method: A new method to handle aberrant responses in psychological and educational testing

LIU Yue[1], LIU Hongyun[2,3]

([1] *Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu* 610066, *China*)

([2] *Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing* 100875, *China*)

([3] *Faculty of Psychology, Beijing Normal University, Beijing* 100875, *China*)

Abstract: The mixture model method (MMM) is a new method proposed to handle data contaminated by aberrant responses in psychological and educational measurement. Compared to the traditional response time threshold methods and the response time residual methods, MMM shows the following advantages: (1) MMM detects aberrant responses and obtaining parameter estimates simultaneously; (2) it precisely recovers the severity of aberrant responding. Through building different item response models and response time models for different latent groups, MMM helps to identify aberrant responses from normal responses. Future researches could investigate the performance of MMM when its assumptions are violated or using data with other types of aberrant response patterns. The computation efficiency of MMM is also likely to be improved by fixing part of the item parameter estimates or by using an optimal way of choosing suitable methods.

Key words: aberrant responses, response time, threshold, residual method, mixture model